Lead Scoring Project

- Pritam S Panchal

INTRODUCTION

- An education company named X Education sells online courses to industry professionals.
- The entity wants to maximize their lead conversion to upto 80%.
- With the use of a Logistic Regression model we will analyze and create a Lead Score to aid the company to contact the most promising leads.

APPROACH & METHODOLOGIES

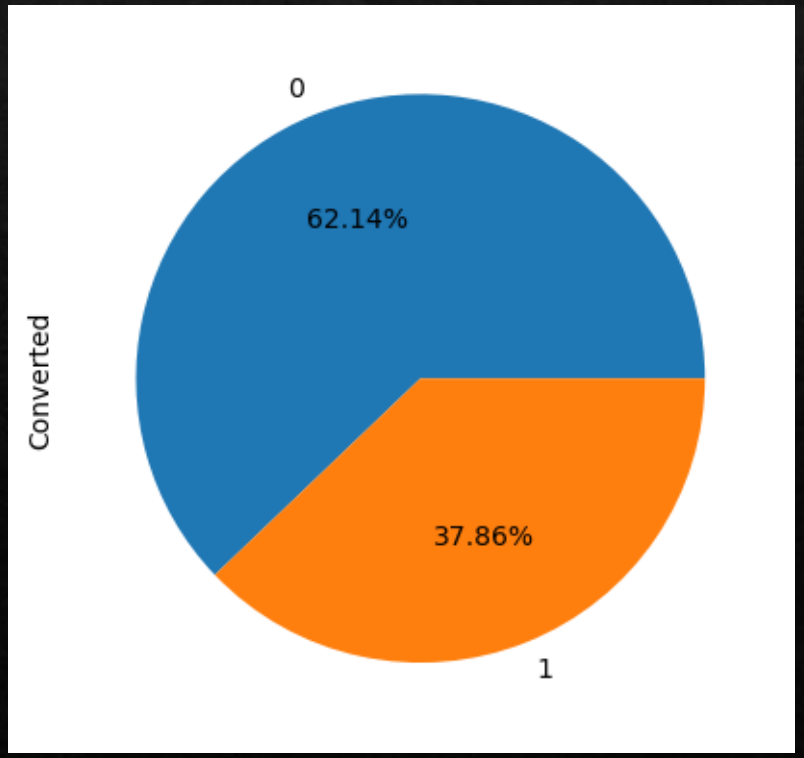
Our approach will consist of the following steps:

- Import the datasets & libraries.
- Check for general structure of the data.
- Handling missing values & Outliers.
- Analyze using Univariate and Bivariate Analysis.
- Creating dummy variables.
- Splitting data and further performing Feature Scaling & Selection.
- Building a Logistic Regression Model
- Model Evaluation
- Making predictions on test set and evaluating again & assigning a Lead Score

GRAPHS & INSIGHTS

Current Lead Conversion

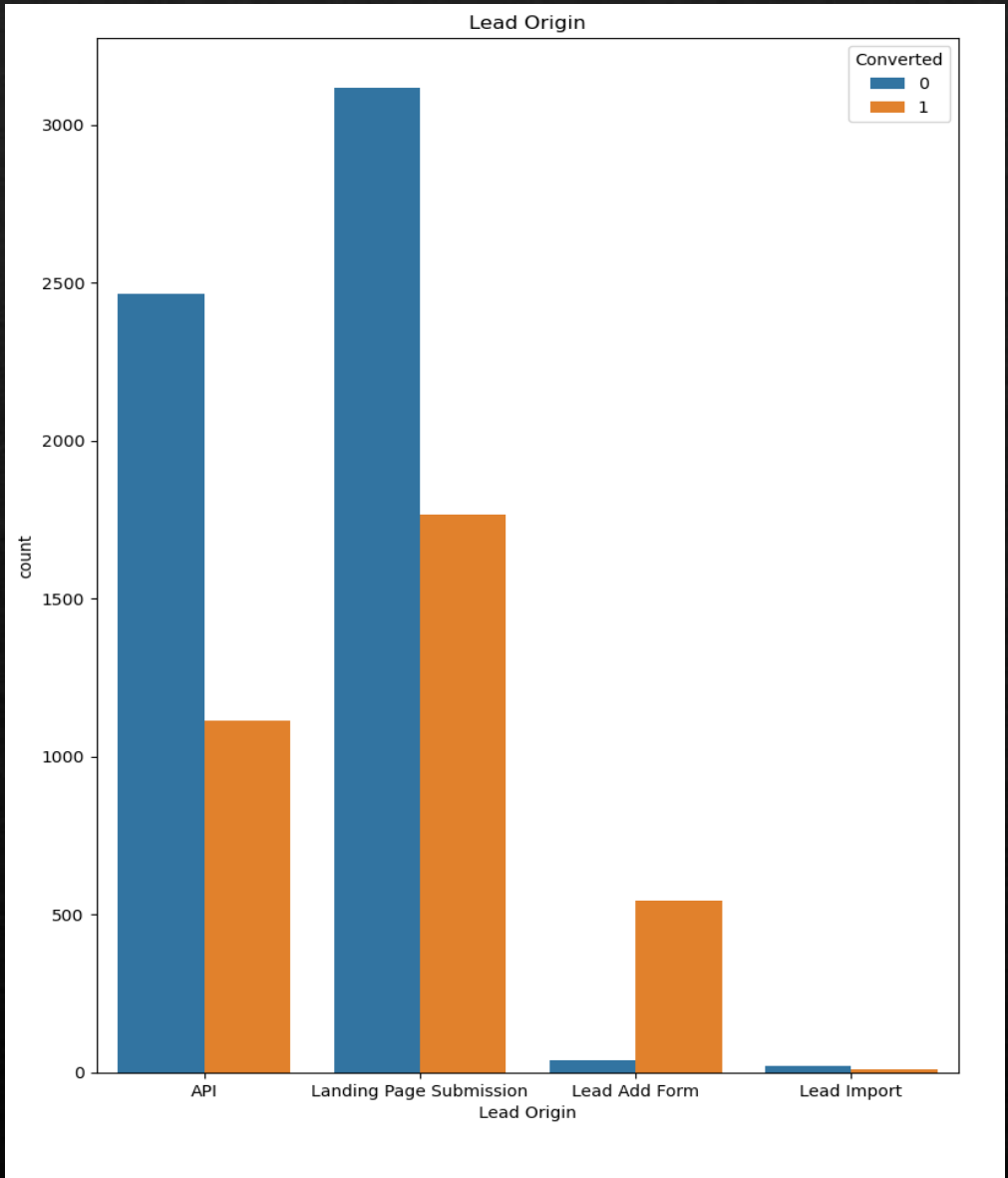
- The current lead conversion rate of X Education was 37.86%.
- We have to boost up the rate above 80%.



Outlier Analysis

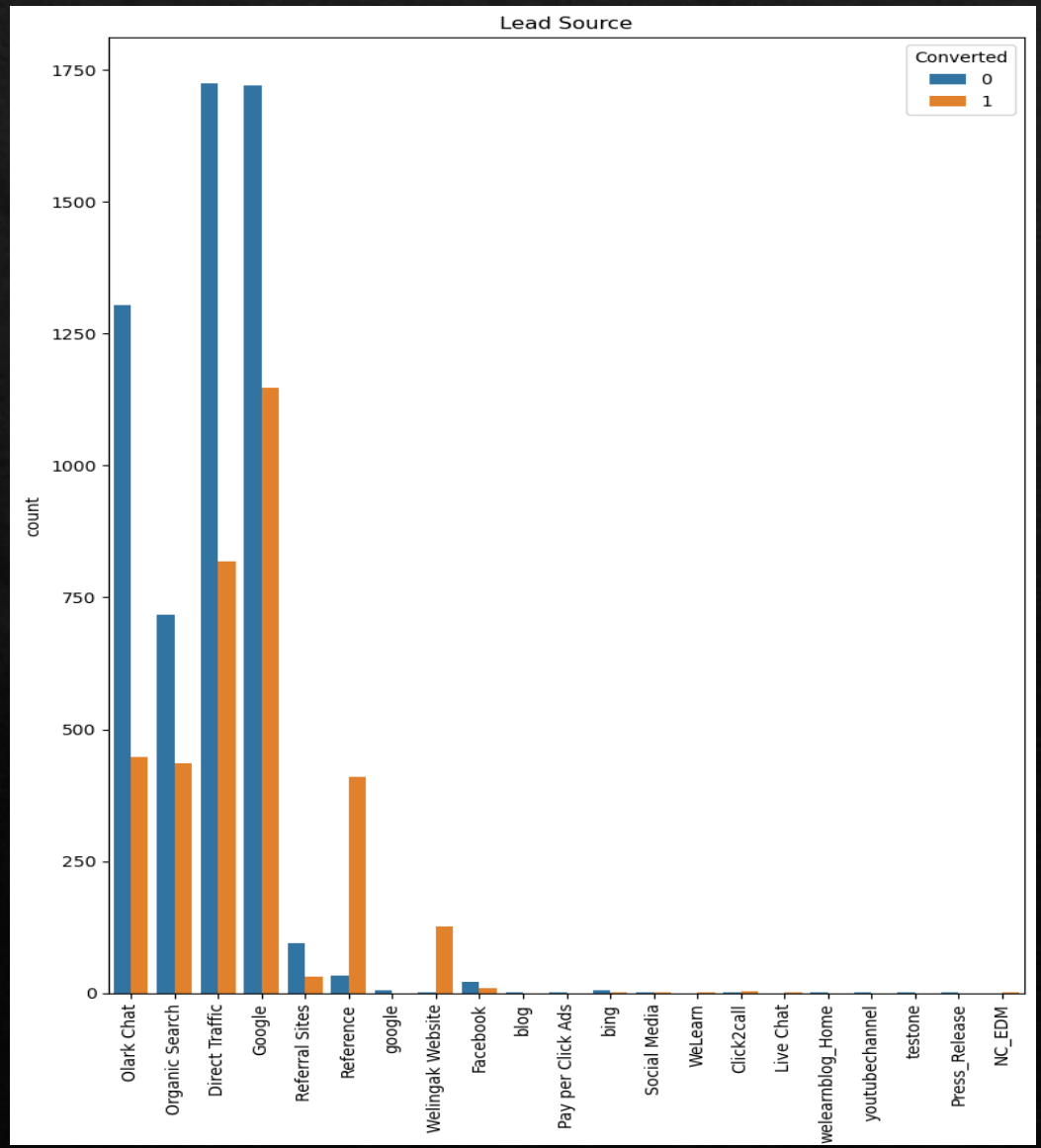
- There were no such major outliers that seem to be impractical

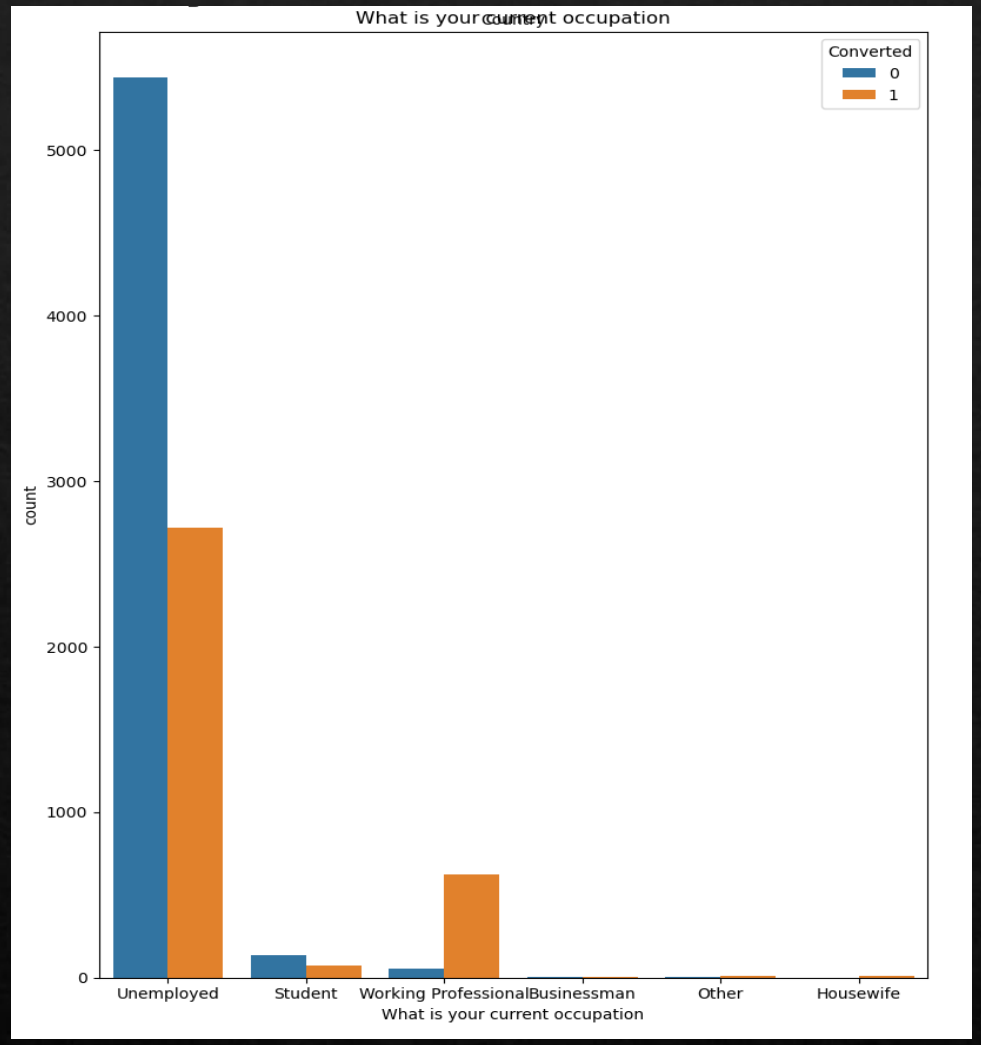
	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
max	251.000000	2272.000000	55.000000



- The conversion from Lead add form is more with respect to others.

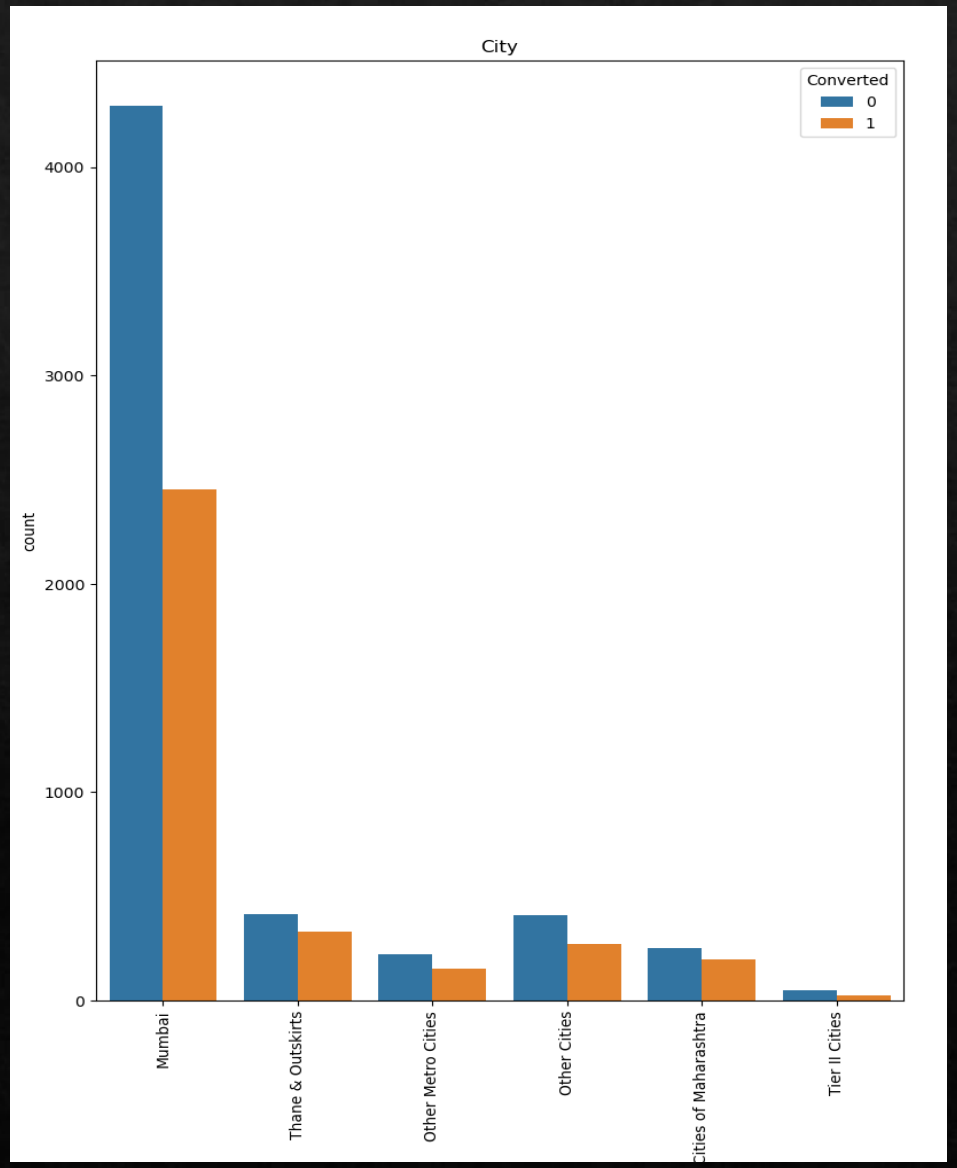
- Maximum number of leads are generated from google and direct traffic.
- The maximum conversion is through reference and welingak website.

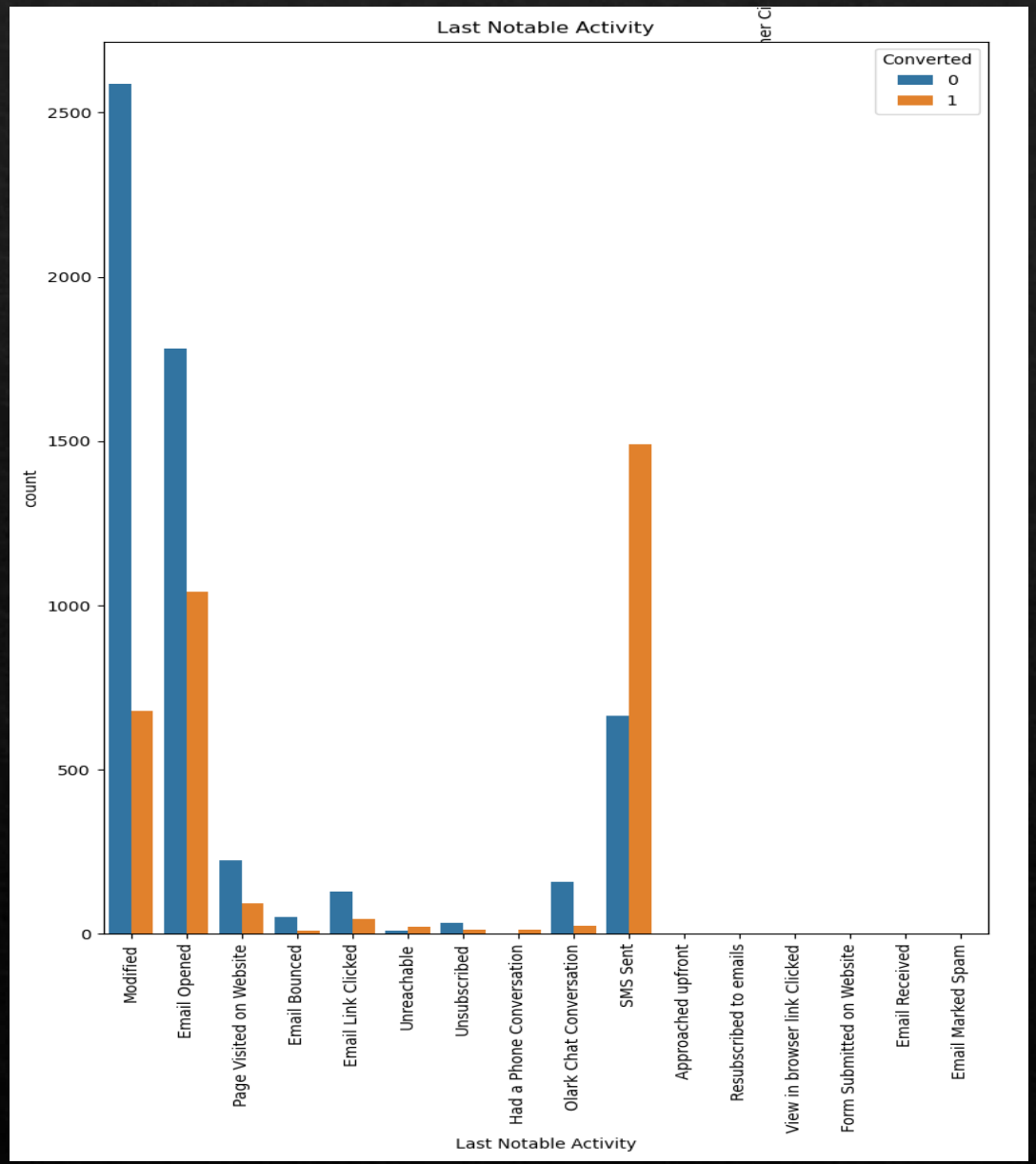




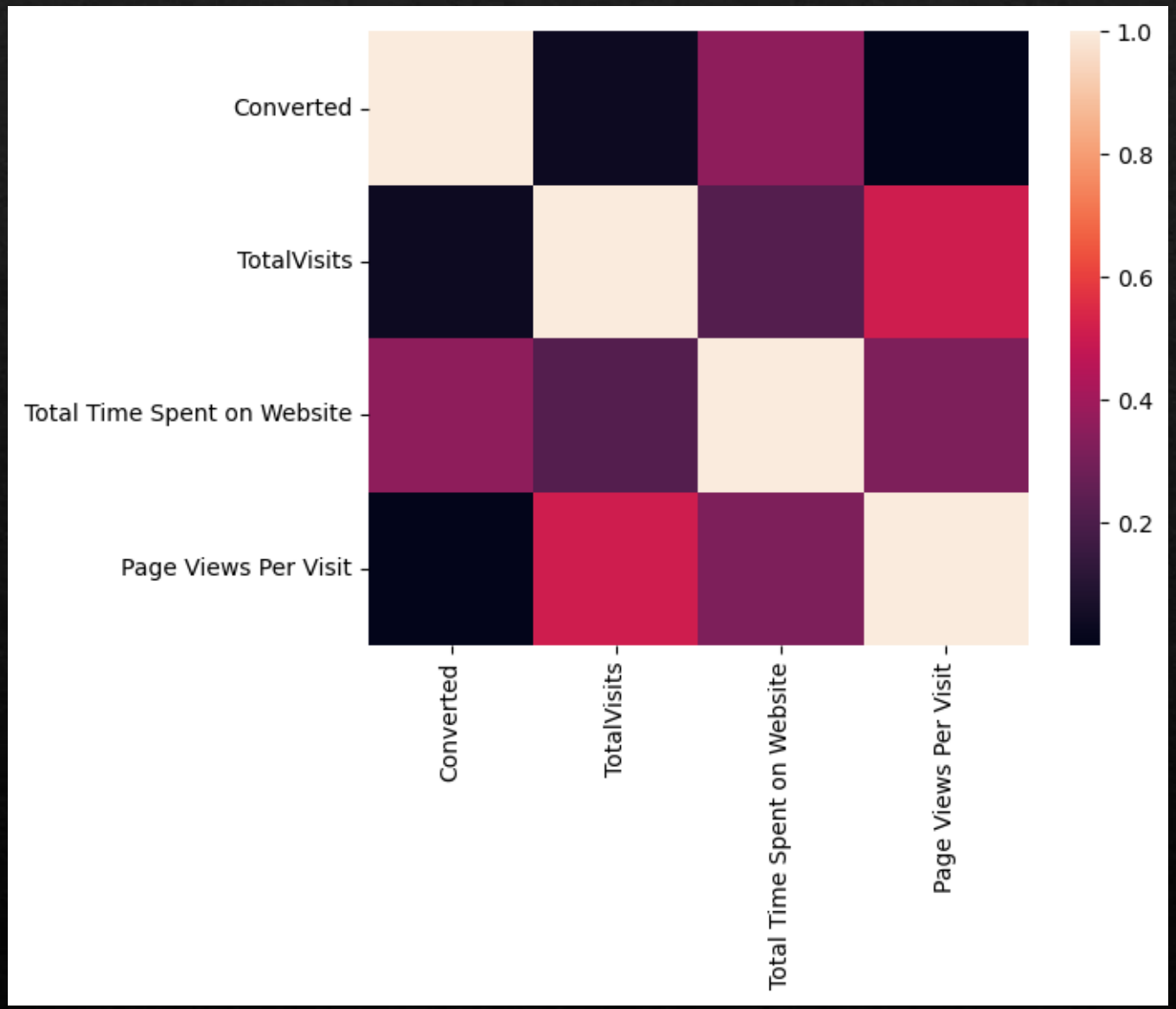
- Most leads are unemployed but maximum conversions are done from working professionals.

- Most leads are from Mumbai, India.





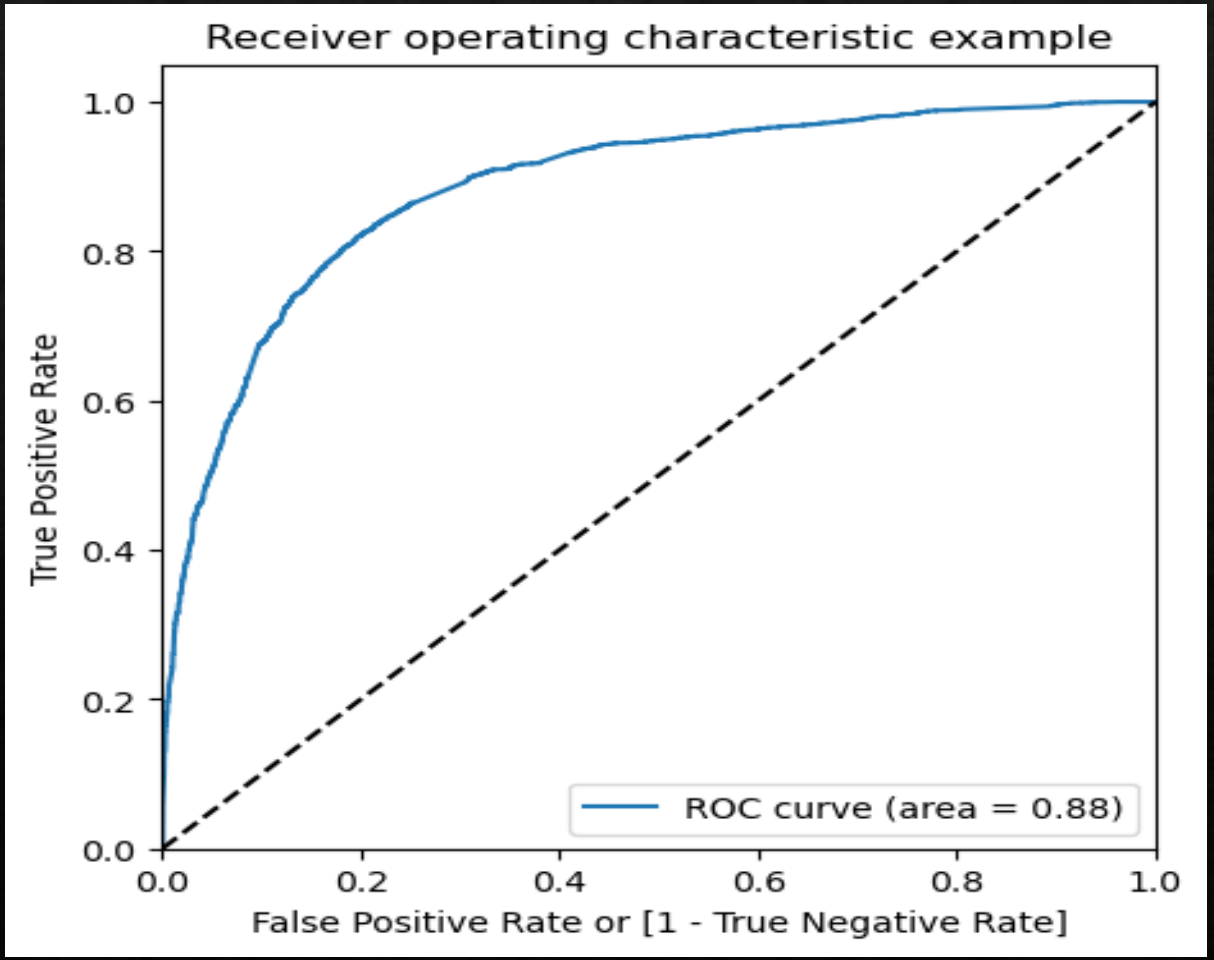
- Leads with SMS Sent as last activity have high conversion rates.



- When Total time spent on website is more, lead conversion is also more.
- The correlation of the other numeric variables is very less wrt. to target variable.

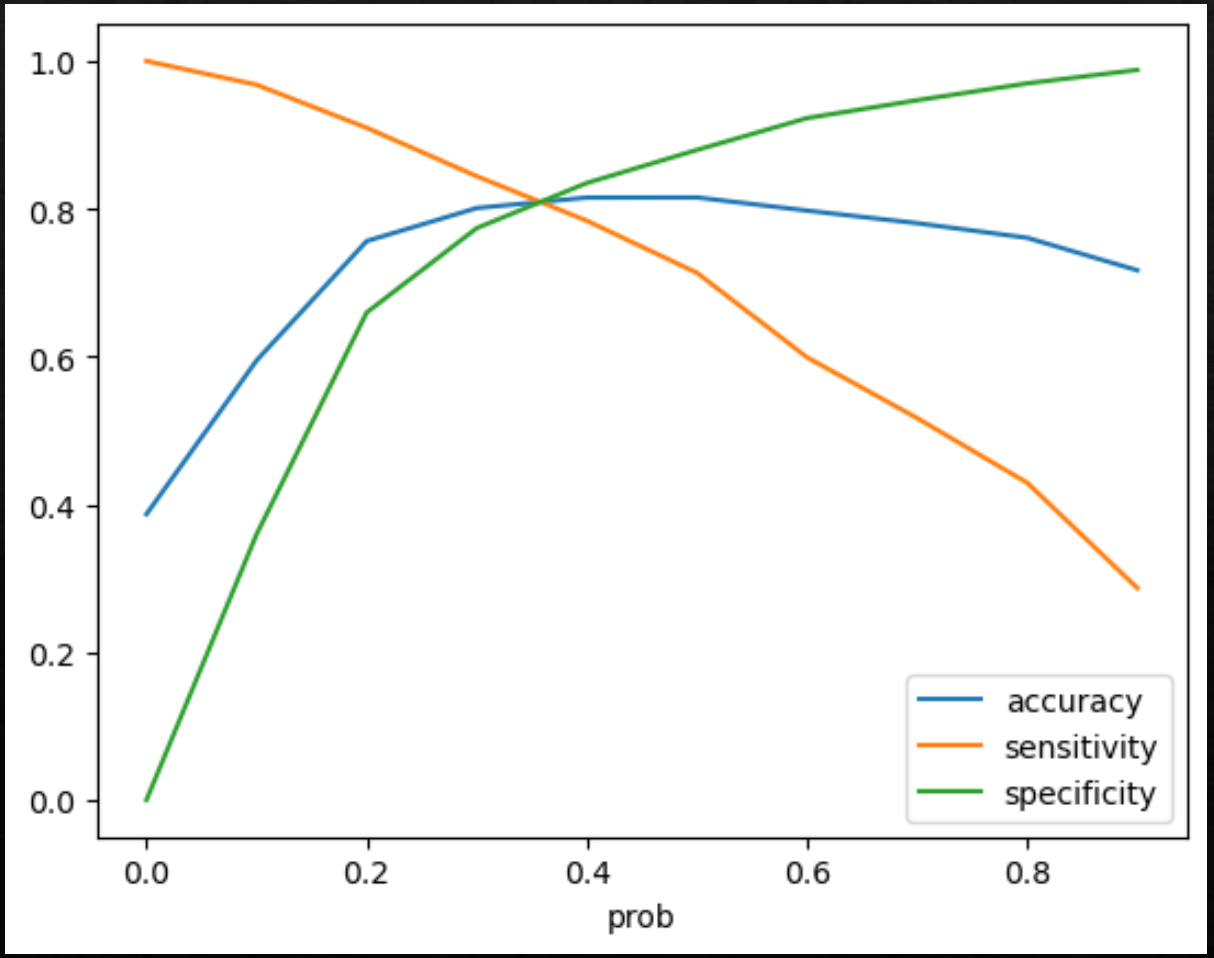
ROC Curve

- Since the ROC curve has more area(0.88) under it and it is closer to the top left corner, our model has turned out to be great.



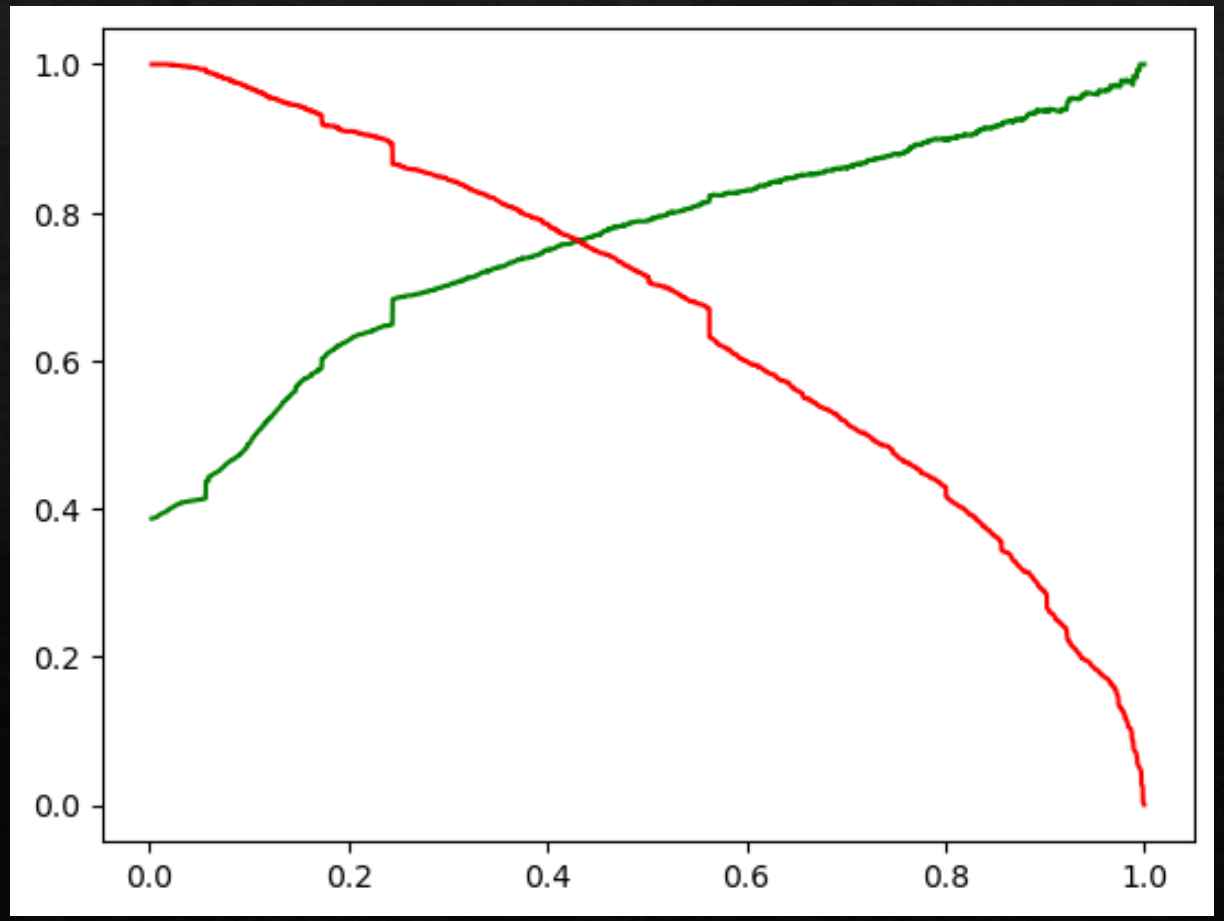
Optimal Value of cutoff

- From the graph we have an optimal cutoff value of 0.35.



Precision Recall Curve

- With the updated cutoff of 0.35, we have precision of around 79% and recall at 71%.



Result

Final Model Parameters:

Train set

- Accuracy : 80.94%
- Sensitivity : 81.55%
- Specificity : 80.56%

Test set

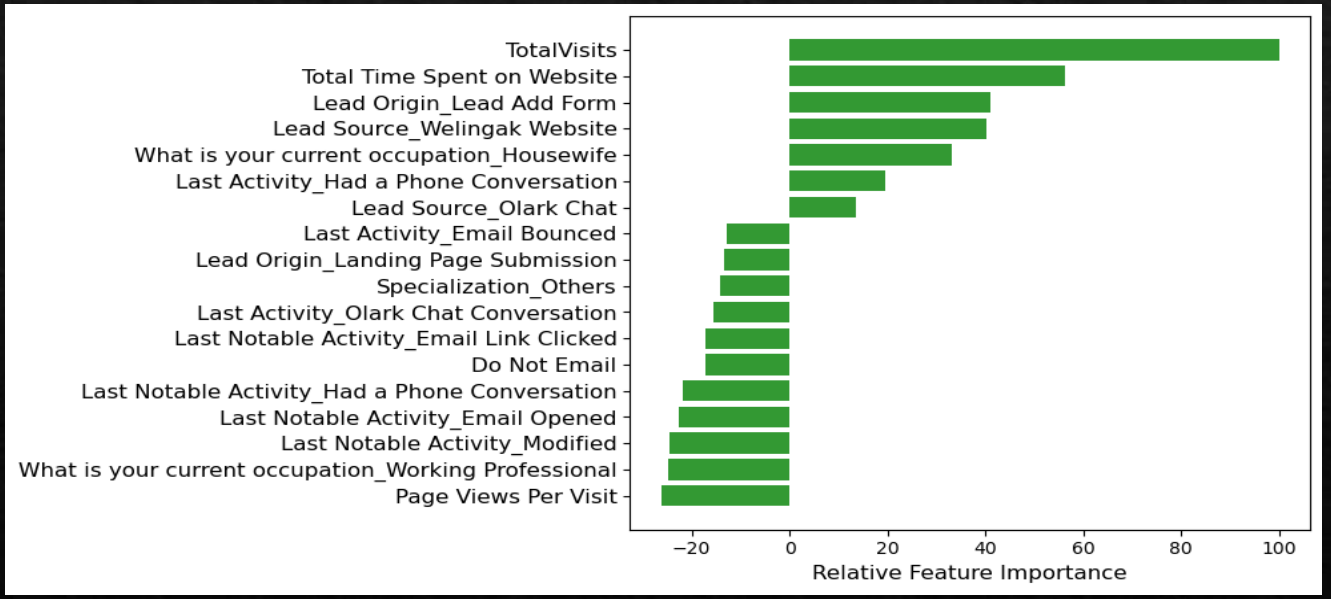
- Accuracy : 80.86%
- Sensitivity : 82.32%
- Specificity : 80.04%

Hence, we have achieved a lead conversion rate of more than 80%.

CONCLUSION

- **Insights:**

After analyzing the datasets we found the variables in the model which contribute most towards the probability of a lead getting converted:



1) TotalVisits
2) Total Time Spent on Website
3) Page Views Per Visit
4) Lead Origin_Lead Add Form
5) Lead Source_Welingak Website
6) What is your current occupation_Working Professional

◈ Suggestions:

1) The team should make phone calls if the lead has visited the website multiple times and also spend more time on it. To increase this the website should be built more intuitive and engaging.

2) Leads from source Welingak Website and having current occupation_Working Professional should be kept a preference.

3) Company should make a cutoff of the lead score and contact only the leads above the cutoff score.

4) To reduce calls, company can send automated emails and SMS system to keep the interaction live with the hot leads.